Supplement to "An improved false discovery rate estimation procedure for shotgun proteomics"

Uri Keich*[1], Attila Kertesz-Farkas[2], and William Stafford Noble*[2,3]

[1]School of Mathematics and Statistics F07, University of Sydney NSW 2006, Australia
[2]Department of Genome Sciences, University of Washington, Foege Building S220B, 3720 15th Ave NE, Seattle, WA 98195-5065
[2]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-5065

1 Supplementary Note 1: The C-TDC and T-TDC procedures are unbiased estimators of false discovery rate

1.1 Elias and Gygi target-decoy competition

If we make the assumption that for each fixed i, Y_i and Z_i are independent and identically distributed given X_i, and in particular $P(Z_i > Y_i | X_i) = 1/2$, then we can prove the following claim establishing the symmetry on which TDC relies. Note that typically we make an even stronger assumption that Y_i and Z_i are independent and identically distributed, as in our mixture model.

Claim 1. If Y, Z are independent and identically distributed given X then, assuming ties are randomly broken by flipping a fair coin,

$$P(Z > Y \mid \mathcal{F}) = P(Z < Y \mid \mathcal{F}) = \frac{1}{2}. \tag{1}$$

Here, $\mathcal{F} = \{\max(Y, Z) > \max(X, T)\}$ corresponds to the event: the PSM between σ and its best match in $db \oplus dc$ is a false positive.

Remark. In order for TDC to work, Elias and Gygi postulate that "incorrect assignments from target or decoy sequences are equally likely" which is essentially this claim. Therefore, you can take this claim as the initial assumption if you feel that the above assumption is too strong. Elias and Gygi also require that "no correct peptides are found in both target and decoy portions" which we can interpret as: for any $\sigma \in \Sigma_1$, $GP(\sigma) \notin dc$. However, because in our setup we randomly break ties we see no real need for the latter constraint.

Proof. Because ties are broken it is clear that

$$P(Z > Y \mid \mathcal{F}) + P(Z < Y \mid \mathcal{F}) = 1.$$

Since Y and Z are iid, the joint distribution function of (X, Y, Z), $F_{X,Y,Z}$, factors as

$$F_{X,Y,Z}(x, y, z) = F_X(x) \cdot G_x(y) \cdot G_x(z),$$

where $G_x(y) = P(Y \le y | X = x)$ is the conditional distribution function of Y (and Z) given that $X = x$. It follows that any two events which are symmetric in Y and Z should have the same probability, and in particular we have

$$P(Z > Y \mid \mathcal{F}) = P(Z < Y \mid \mathcal{F}).$$

*Correspondence to uri@maths.usyd.edu.au, Phone: 61 2 9351 2307 and william-noble@uw.edu, Phone: 1 206 221-4973, Fax: 1 206 685-7301

The claim now follows. □

Note that Y_i need not be identically distributed for this claim to hold, however, we do need to assume that given X_i, Y_i is independent of Z_i and the two have the same distribution. Alternatively we can assume that the corollary of the latter claim, (1), holds. Whether or not this condition is approximately satisfied in practice depends on the definitions of the target and decoy databases as well as on the scoring scheme.

Is C-TDC biased? The answer to this question depends how you look at it.

Claim 2. Under the assumptions of Claim 1

$$E(F_C) = 2E(F_D) = E(\widehat{F_C})$$

Proof. Let $F_T = |\{i : Y_i > \max(x_i, Z_i, T)\}|$ be the unobserved number of false discoveries that fall in the target *db*. Multiplying (1) for each i by $P(\mathcal{F}_i)$ (or, alternatively invoking again the symmetry in Y and Z) we have, for all i

$$P\left(Z_i > \max\left\{X_i, Y_i, T\right\}\right) = P\left(\left\{Z_i > Y_i\right\} \cap \mathcal{F}_i\right) = P\left(\left\{Z_i < Y_i\right\} \cap \mathcal{F}_i\right) = P\left(Y_i > \max\left\{X_i, Z_i, T\right\}\right). \quad (2)$$

Summing over all i we find that

$$E\left(F_D\right) = \sum_i P\left(Z_i > \max\left\{X_i, Y_i, T\right\}\right) = \sum_i P\left(Y_i > \max\left\{X_i, Z_i, T\right\}\right) = E\left(F_T\right). \quad (3)$$

The claim now follows from $F_C = F_T + F_D$. □

The last claim shows that $\widehat{F_C}$ is an unbiased estimator of F_C, which is the number of false positives in the concatenated search against the combined dataset $db \oplus dc$. Moreover, if n_Σ is rather large and does not contain "too many" repetitive spectra then with high probability $F_D/F_C \approx 1/2$; therefore, $2F_D/D_C$ is a reasonably good estimator of the rate of false discoveries in the concatenated search.

However, recall that we are interested in the rate of false discoveries in the set of discoveries that fall within the target database. The latter has only $D_T = D_C - F_D$ PSMs in it, and the number of false discoveries among those is

$$F_T = \left|\{i : Y_i > \max(X_i, Z_i, T)\}\right|. \quad (4)$$

Therefore, the FDR among our reported discoveries is

$$\frac{F_T}{D_T} = \frac{F_T}{D_C - F_D} \leq \frac{F_T + F_D}{D_C},$$

where the latter inequality, which follows from $F_T \leq D_T$, is strict when $F_D > 0$ (which essentially always occurs in a reasonable search). Note that the right-hand side of the last equation is the FDR in the concatenated search, which is what the C-TDC estimator of $2F_D/D_C$ is trying to predict. It follows that C-TDC is conservatively biased when it comes to estimating the FDR in the set of target discoveries.

Remark. Although this was not the original intent of Elias and Gygi, the next claim shows that the C-TDC estimator is sufficiently conservative to allow us to conservatively estimate the FDR among the set of D discoveries when searching the target database *on its own*. Indeed, the claim proves what is intuitively obvious: F_C/D_C is larger than F/D, since the concatenated database allows more false discoveries without increasing the number of true discoveries.

Claim 3.

$$F_C/D_C \geq F/D \quad (5)$$

Proof. Let $N_f = F_C - F$ and similarly $N_d = D_C - D$. It is easy to see that $N_f \geq N_d \geq 0$, and since $F_C/D_C = \frac{F+N_f}{D+N_d}$ and $F \leq D$ the claim follows. □

Note that the inequality in (5) will typically be a strict one because $N_f > N_d$ for most reasonable setups.

To summarize, C-TDC presents an asymptotically unbiased estimator of F_C/D_C, the FDR in the concatenated database. However, it is a conservative estimator of F/D, the FDR in the target-only set of discoveries, as well as of F_T/D_T, the FDR in the target-filtered set of TDC discoveries. This result means that if we want to control the FDR at level α using C-TDC we will need to set the threshold T higher than we would have had we been able to directly control F/D at the same level α. In practical terms, this means that C-TDC is conservative when compared to an ideal tool that allows us to control F/D directly.

1.2 "Target-only" target-decoy competition

First note that D_T is observable, and we showed that $E(F_D) = E(F_T)$ (3) so T-TDC has the right expectation in some sense. Moreover, for an independent set of spectra $F_T/F_C \longrightarrow 1/2$ with probability 1 as $n_\Sigma \to \infty$ (we are assuming here that $F_C \longrightarrow \infty$ as well, as will be the case for any reasonable setup), and it follows from $F_C = F_T + F_D$ that $F_T/F_D \longrightarrow 1$ with probability 1. Therefore, if the set of spectra is large and "fairly independent" we expect that T-TDC $\approx F_T/D_T$. More precisely, the ratio between the two quantities goes to 1 almost surely when the size of the set of independent spectra goes to infinity:

Claim 4. If the spectra are independent, and the number of spectra $n_\Sigma \longrightarrow \infty$ in such a way that the number of false discoveries $F_C \longrightarrow \infty$ then with probability 1

$$\frac{\text{T-TDC}}{F_T/D_T} \longrightarrow 1.$$

Remark. We again stress that in the TDC approach we do not have to assume that the Y_i (or Z_i) are identically distributed, rather that separately for each fixed i, *given* X_i, Y_i and Z_i are independent and identically distributed.

2 Supplementary Note 2: STDS-PIT underestimates the true FDR

While STDS-PIT provides an intuitively appealing way to estimate the FDR, we argue that in the context of our model this method underestimates the FDR. The problem goes back to π_0, the proportion of true null hypotheses. In their paper, Käll et al. call this proportion *PIT* (percentage of incorrect targets), suggesting that the PIT estimates the overall rate of incorrectly identified PSMs in the target database. However, it is important to understand what is the null hypothesis according to which the p-values are computed here. Because the p-values are estimated from the decoy set, the null hypothesis is that the spectrum is foreign. Therefore, if we use the FDR analysis of Storey[1], then the resulting $\widehat{\pi_0}$ estimates $|\Sigma_0| / |\Sigma| = 1 - \pi_1$, the proportion of foreign spectra in our input set. Of course, this estimate ignores all the incorrect targets that are attributed to the native spectra.

One might argue that Käll et al. define the problem differently. Indeed, their setup is such that they only care whether the PSM is random (null hypothesis) or a correct one (alternative hypothesis). Using this setup the proportion of true null hypotheses as estimated by Storey's FDR analysis does indeed coincide with their desired PIT, thus seemingly validating their estimate.

However, note that the p-values to which the FDR analysis is applied are estimated from a set of foreign spectra (all spectra are foreign relative to the decoy database). Therefore, if you accept the model presented here, then these p-values are not always correctly calculated relative to the null "the PSM is random". Indeed, in relying on the decoy set to estimate the p-value of a PSM involving a native spectrum $\sigma_i \in \Sigma_1$ we ignore the fact that the optimal random match y_i needs to compete with the score of the true match x_i. For example, no random PSM involving σ_i can score lower than x_i; therefore, it could not be assigned a p-value less significant than that assigned to x_i. This is in contradiction to the fact that p-values of true nulls should be distributed uniformly on [0,1]. Thus, the p-values of PSMs involving $\sigma \in \Sigma_1$ are overestimated (too small), which in turn leads to underestimating the true PIT (too small).

3 Supplementary Note 3: suggested implementation of mix-max

Equation (9) in the main paper shows how we estimate the FDR in the target list of discoveries defined by all PSMs whose score exceeds the threshold T. In practice the user is typically interested in estimating the FDR associated with accepting the top k target PSMs for all possible values of $k = 1, 2, \ldots, n_\Sigma$.

More precisely, we want the estimated FDR assuming we accept all target PSMs with score $s \geq w_m$ for each w_m in the set of all target PSM scores $\{w_j : j = 1, \ldots, n_\Sigma\}$. For any fixed w_m this can be done using the main paper (9) by choosing any threshold T which satisfies

$$\max\{z_j : z_j < w_m\} < T < w_m. \tag{6}$$

Note that the target discovery list associated with any T which satisfies (6) is the set $\{w_j : w_j \geq w_m\}$. Similarly, it is easy to see that for any such T the mix-max estimated FDR (9) in the main paper is equivalent to

$$\widehat{\frac{F}{D}}(w_m) := \frac{\widehat{\pi_0} \cdot \sum_{j=1}^{n_\Sigma} 1_{z_j \geq w_m} + (1 - \widehat{\pi_0}) \cdot \sum_{z_j \geq w_m} \left[\frac{\sum_k 1_{w_k \leq z_j}}{(1-\widehat{\pi_0}) \cdot \sum_k 1_{z_k \leq z_j}} - \frac{\widehat{\pi_0}}{1-\widehat{\pi_0}} \right]_{[0,1]}}{\sum_i 1_{w_i \geq w_m}}. \tag{7}$$

Thus, our goal here is to compute for all target PSM scores $\{w_m\}$ the mix-max estimate of the FDR in (7) above. Rather than calculating the right hand side of (7) separately for each w_m we provide pseudocode that makes more efficient use of the relations between the number of discoveries and estimated false discoveries as we progress along the *sorted list of increasing* target scores $\{w_{(m)}\}$. To take advantage of these relations we introduce in the pseudocode below the variables

$$N_{w \leq z}(j) = \sum_k 1_{w_k \leq z_{(j)}} \quad N_{z \leq z}(j) = \sum_k 1_{z_k \leq z_{(j)}} \quad N_{z \geq w}(j) = \sum_k 1_{z_k \geq w_{(j)}} \quad N_{w \geq w}(j) = \sum_k 1_{w_k \geq w_{(j)}}, \tag{8}$$

where $z_{(j)}$ denote the list of decoy scores in increasing order.

Note that in the absence of ties $N_{z \leq z}(j) = j$ but given that we allow ties we only know that $N_{z \leq z}(j) \geq j$. Regardless, since the input to `mixmaxFDR` below is the sorted lists of target and decoy PSM scores ($\{w_{(m)}\}$ and $\{z_{(m)}\}$) the above vectors can readily be computed in time $O(n_\Sigma)$ (in fact, the computation of $N_{z \geq w}$ and $N_{w \geq w}$ can be rolled into the main loop of `mixmaxFDR`).

Algorithm 1 Procedure for estimating the FDR using mix-max. The inputs are the sorted lists of target and decoy calibrated PSM scores, $\boldsymbol{w}_{(i)}$ and $\boldsymbol{z}_{(i)}$, respectively. The output is a list $\boldsymbol{R}$, where $\boldsymbol{R}_m$ is the mix-max estimated FDR of (7) with $w_m = w_{(m)}$.

procedure MIXMAXFDR($\boldsymbol{w}_{(i)}$, $\boldsymbol{z}_{(i)}$)
  Estimate $\boldsymbol{p}$, the p-values of $\boldsymbol{w}$, using the empirical distribution of $\boldsymbol{z}$
  Compute $\widehat{\pi_0}$ by applying `R`'s `qvalue` function to $\boldsymbol{p}$
  $n \leftarrow \text{length}(\boldsymbol{w})$
  if $\widehat{\pi_0} = 1$ **then**
    Return a vector of n 1s
  end if
  Compute the vectors $N_{w \le z}$, $N_{z \le z}$, $N_{z \ge w}$, $N_{w \ge w}$ ▷ See (8) above
  $E_1 \leftarrow 0$
  $j \leftarrow n$
  for $m \leftarrow n \ldots 1$ **do**
    while $(j > 0)$ & $(\boldsymbol{z}_{(j)} \ge \boldsymbol{w}_{(m)})$ **do**
      $\hat{p} \leftarrow [N_{w \le z}(j) - \widehat{\pi_0} \cdot N_{z \le z}(j)] \,/\, [(1 - \widehat{\pi_0}) \cdot N_{z \le z}(j)]$
      $E_1 \leftarrow E_1 + \min\{1, \max\{0, \hat{p}\}\} \cdot (1 - \widehat{\pi_0})$
      $j \leftarrow j - 1$
    end while
    $\boldsymbol{R}(m) \leftarrow [\widehat{\pi_0} \cdot N_{z \ge w}(m) + E_1] \,/ N_{w \ge w}(m)$
    $\boldsymbol{R}(m) \leftarrow \min\{\boldsymbol{R}(m), 1\}$
  end for
  return $\boldsymbol{R}$
end procedure

4 Supplementary Note 4: the effects of our 10K-calibration procedure on FDR estimation

To test whether the way we calibrate our scores using 10K decoys (see Methods) significantly impacts any of the FDR estimation methods, we slightly modified our simulated data experiment. Specifically, we added a calibration step whereby each randomly drawn PSM score was calibrated using a "spectrum-specific" sample of 10K scores drawn according to the null distribution. Note that our *simulated* PSM scores were perfectly calibrated to begin with, so this redundant calibration procedure only compromised the initial perfect calibration. Its goal, however, was to mimic the effect of our spectrum-specific calibration procedure, which is applied to the real data where perfect calibration cannot be achieved.

Looking at the accuracy of the estimated FDR (Supplementary Figure 7) we note that overall it is quite similar to the accuracy on the original simulated data (main Figure 3) with the notable exception that both T-TDC and mix-max slightly overestimate the FDR (too conservative) for moderately large spectrum sets ($\geq$10K) and small FDR value ($<$0.01).

The reason for T-TDC changing its behavior from underestimating to overestimating the FDR for larger spectrum sets has to do with the fact that once the set of spectra becomes as large as the number of decoys we use in our non-parametric calibration it is very difficult to avoid false discoveries with maximal possible score. These false discoveries in turn tend to inflate the estimation of small FDR values. For example, imagine that a set of 30K spectra, of which 15K are native, has 10% true PSMs that receive a maximal score under the 10K calibration procedure. In this case, the expected number of maximally scoring decoy PSMs is 3 (we are using 10K decoys to calibrate each spectrum score); therefore, on average the smallest T-TDC estimated FDR will be about $3/1500 \approx 0.002$. In all those cases, where an FDR level of 0.001 is not attainable, the actual FDR at that threshold of 0.001 will be 0 because there are no discoveries at that level. (Note that we are using the convention that the FDR is defined as 0 when there are no discoveries.) Hence, in such cases, the actual FDR will be lower than the nominal one, effectively implying that T-TDC is overestimating the true FDR.

Finally, in terms of the relative number of discoveries we see there is little difference compared with the original data (Supplementary Figure 8).

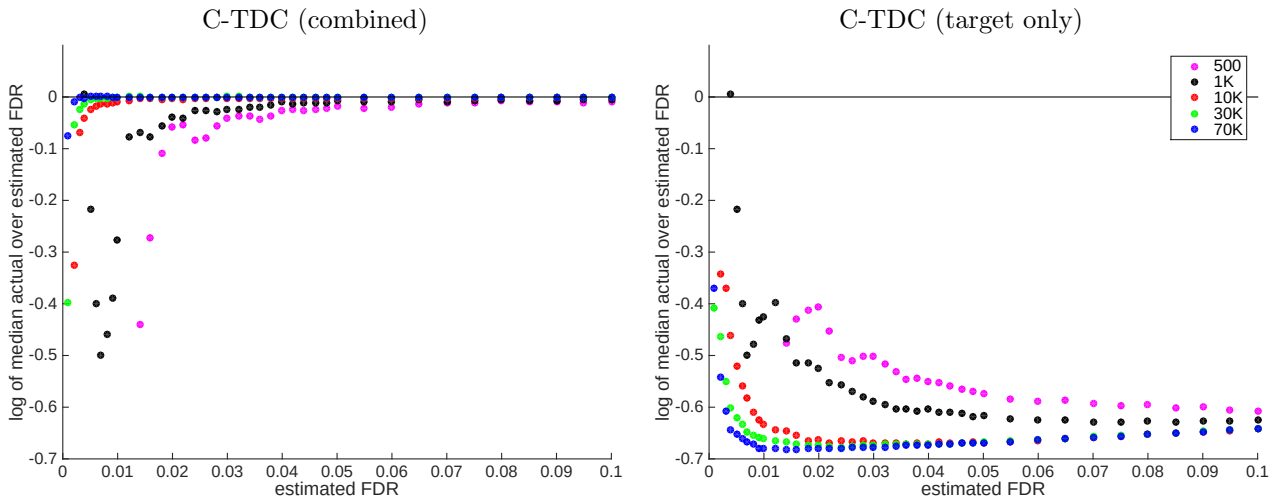

Figure 1: **Accuracy of C-TDC estimated FDR (mixture model).** Similarly to Figure 3 in the main paper, the accuracy of the C-TDC FDR estimation method is gauged by plotting the logarithm of the median ratio between the actual FDR and the nominal one. The graphs were generated using the same data that was used in the said figure. The left figure shows that C-TDC is reasonably well estimating the FDR in the *combined* target and decoy list when the spectra is reasonably large or when the FDR is not too small (say, above 5%). However, if the C-TDC is used to estimate the FDR in the target-only list then as expected it is significantly conservative: the actual FDR in the target list of discoveries is substantially smaller than the estimated one.

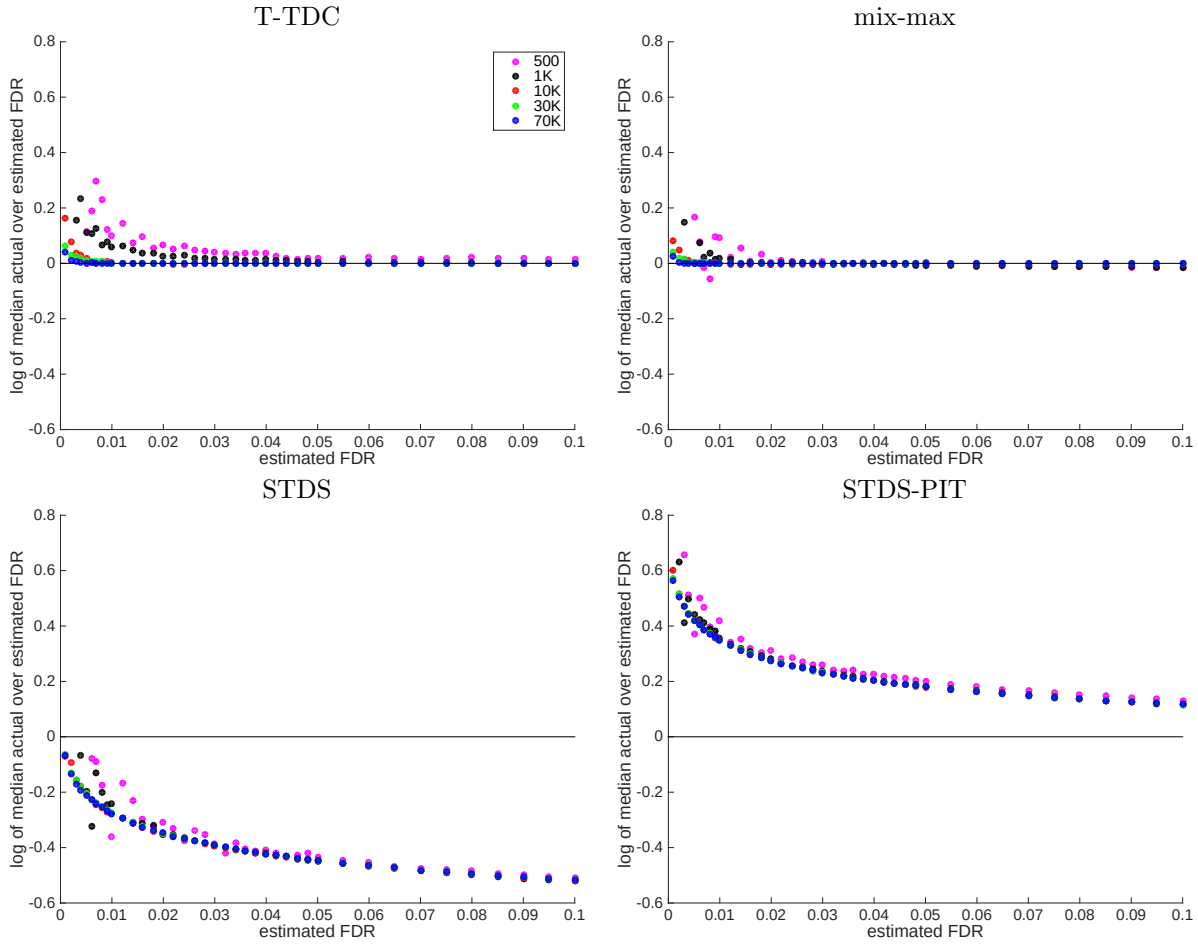

Figure 2: **Accuracy of estimated FDR (mixture model, larger separation).** Same as Figure 3 in the main paper, except the scores of the native spectra matches X_i were drawn from a $N(3.0, 1)$ distribution which increased the average separation between the correct PSMs and the false ones which were still drawn from a $N(0, 1)$ distribution. The results are qualitatively quite similar to the original figure though STDS-PIT is a little less liberal than before while STDS is a little more conservative and both mix-max and T-TDC are overall slightly more accurate.

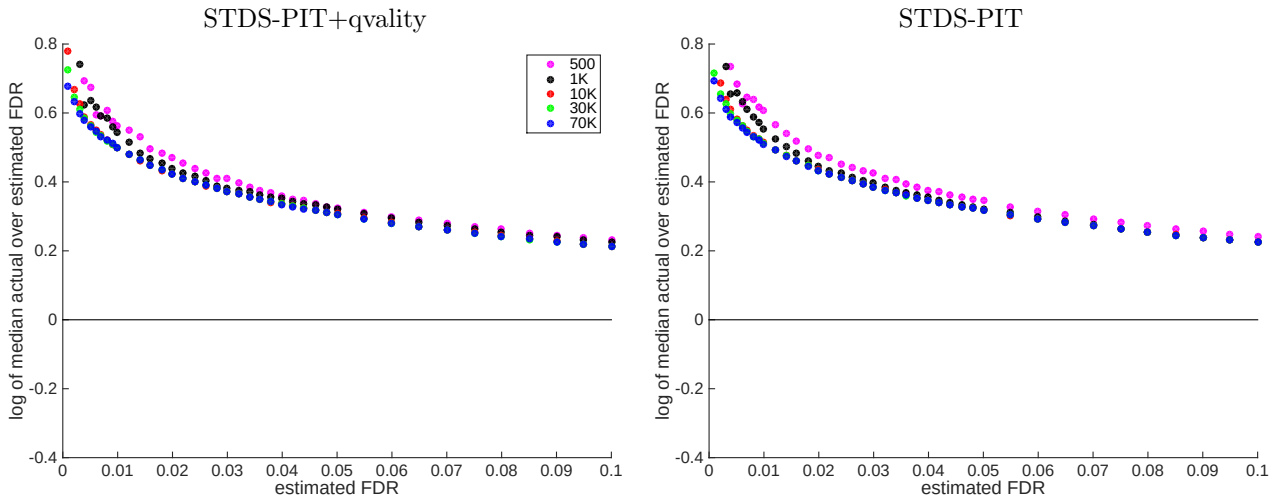

Figure 3: **Accuracy of STDS-PIT+qvality.** Similarly to Figure 3 in the main paper, the accuracy of the STDS-PIT+qvality FDR estimation method is gauged by plotting the logarithm of the median ratio between the actual FDR and the nominal one. The graphs of STDS-PIT+qvality were generated using data that was generated according to the exact same protocol used to generate the data that was used in the graphs of STDS-PIT. The latter is added here so as to ease the comparison between the two methods. It seems that in this setup there is little difference between these methods.

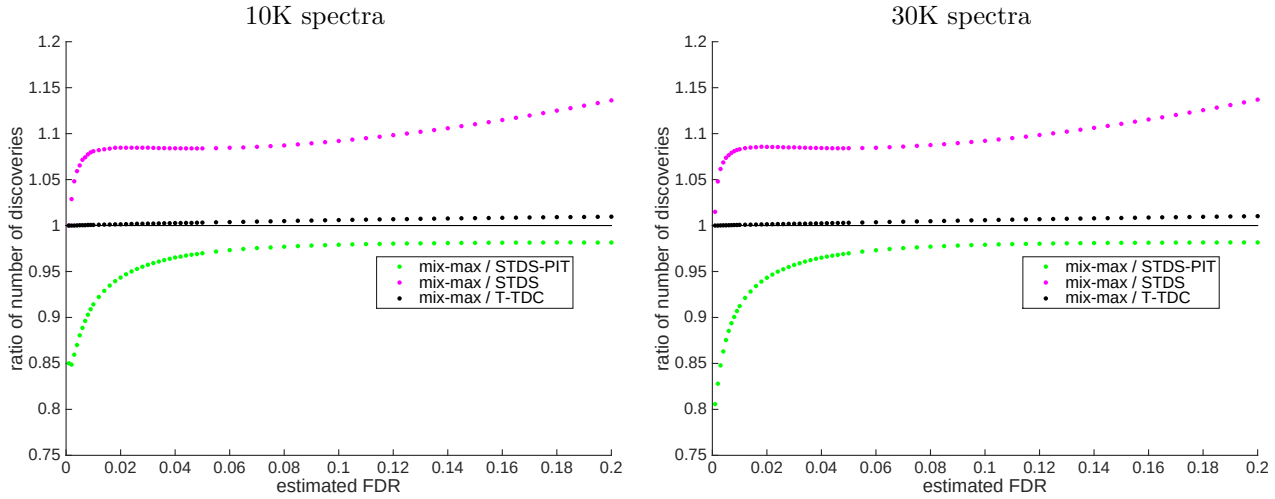

Figure 4: **Median ratios of number of discoveries, larger separation.** Same as Figure 4 in the main text, except the native spectra matches X_i were drawn from a $N(3.0, 1)$ distribution rather than from a $N(2.5, 1)$ distribution which increased the average separation between the correct PSMs and the false ones (which were still drawn from a $N(0, 1)$ distribution). The results are qualitatively quite similar to the original figure although the differences between the methods (and particularly between STDS-PIT and mix-max) naturally diminish as the separation between the native and null scores grows larger. The graphs are again derived from 10K draws.

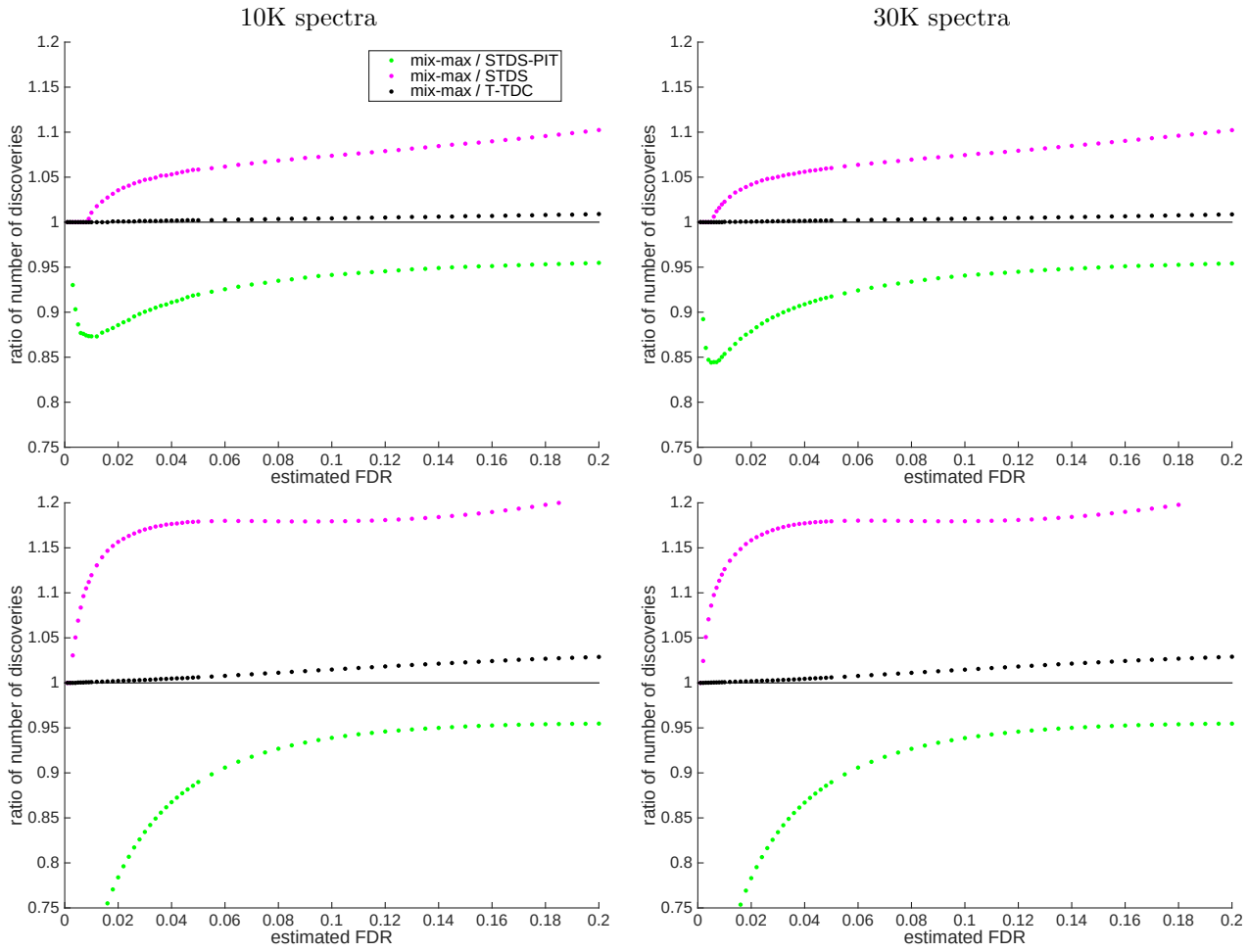

Figure 5: **Median ratios of number of discoveries.** Same as Figure 4 in the main paper except the native spectra rate was set to $1 - \pi_0 = 0.3$ for the two top panels and to $1 - \pi_0 = 0.7$ for the two bottom panels. Comparing these two cases with $1 - \pi_0 = 0.5$ in Figure 4 (main paper) we see that the difference between the methods diminishes with the decrease in the proportion of native spectra.

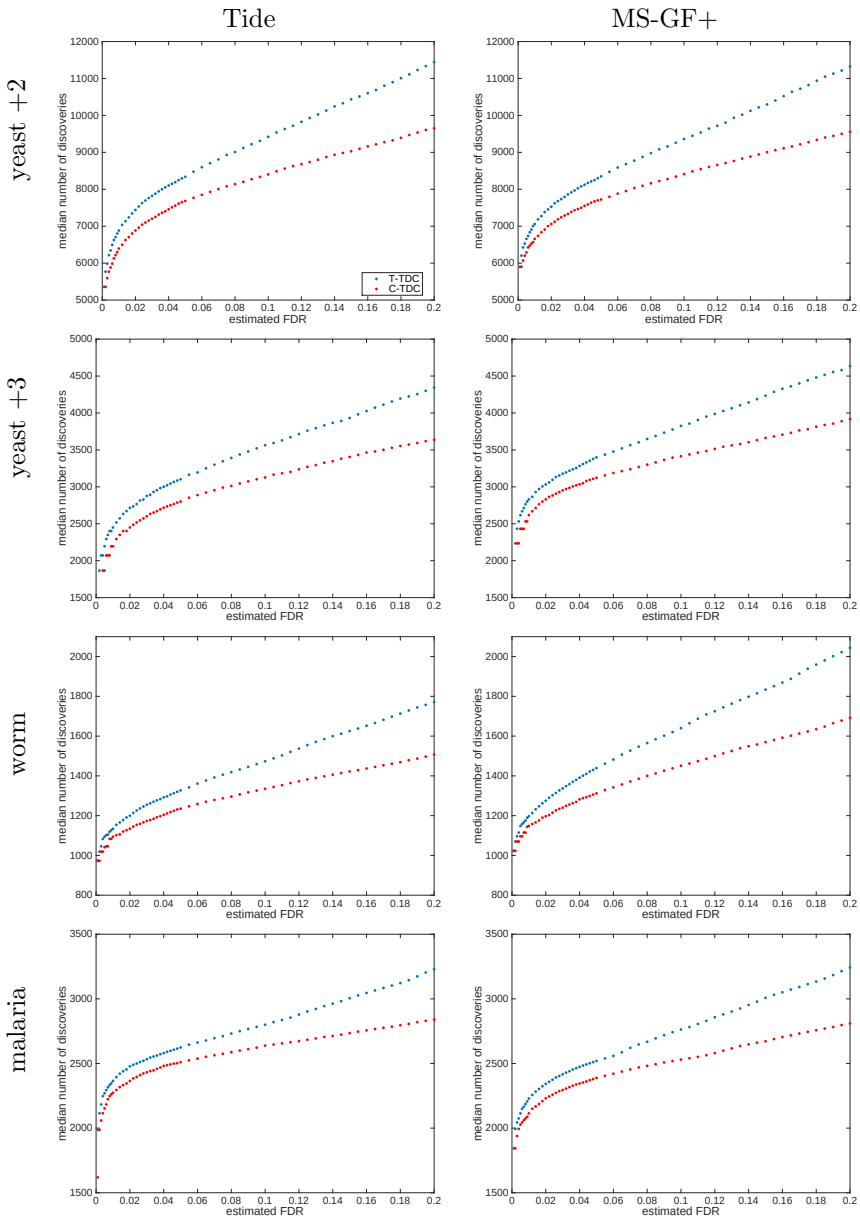

Figure 6: **Median number of T-TDC and C-TDC discoveries in the yeast dataset.** Using the same data as in Figure 6 in the main paper each panel plots, as a function of estimated FDR, the median number of T-TDC and C-TDC target discoveries. The spectrum sets are the yeast, worm and malaria data sets. In each plot, the medians were taken with respect to 1000 corresponding discovery numbers using that many randomly drawn decoy databases. Each pair of target-decoy databases was searched using two different search engines: Tide and MS-GF+. Because C-TDC is estimating the FDR in the combined list of discoveries it is clearly conservative when it is used to determine the number of target discoveries.

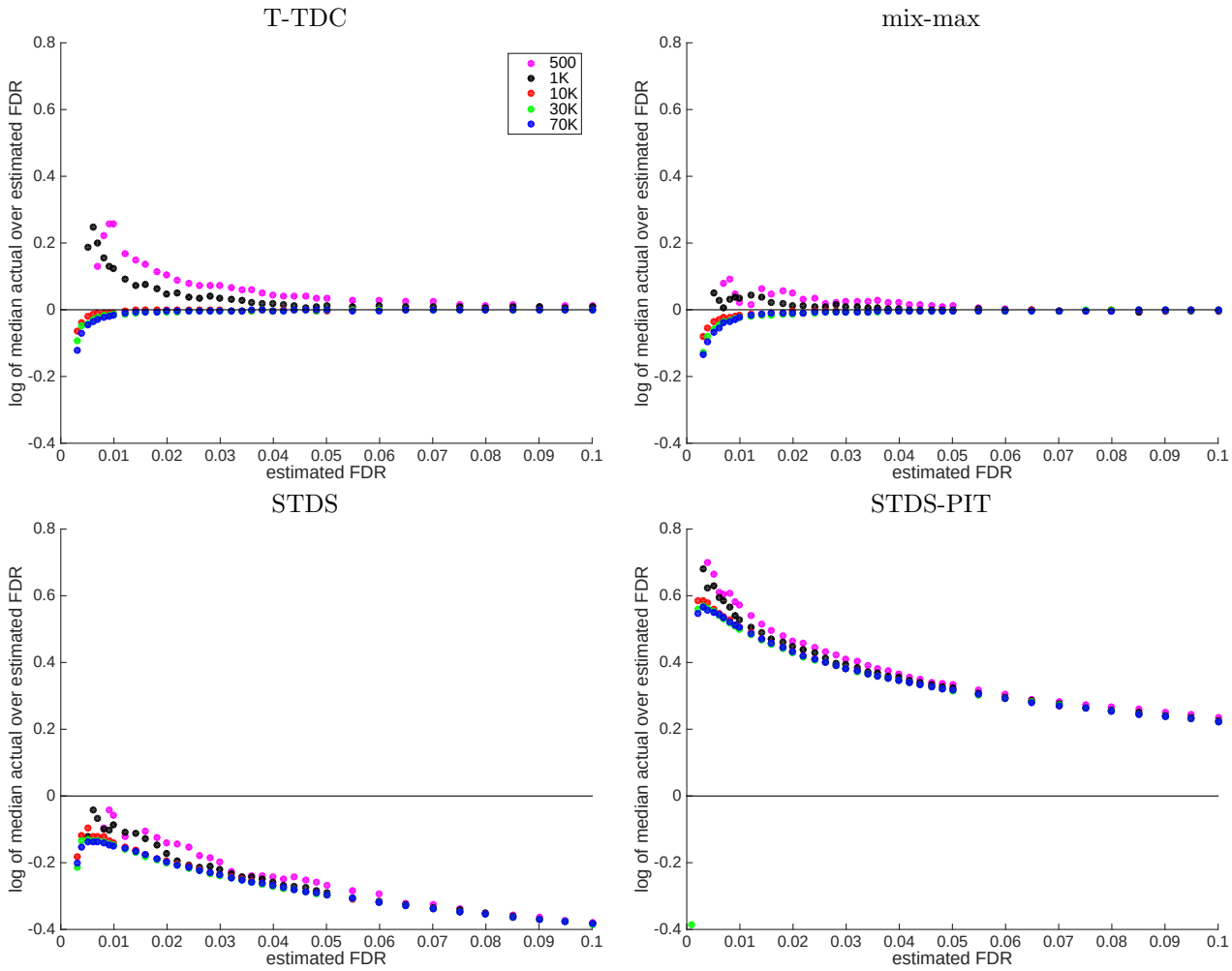

Figure 7: **Accuracy of estimated FDR (mixture model, 10K calibrated scores).** Same as Figure 3 in the main paper, except all randomly drawn PSM scores (correct and false) were first "calibrated" using spectrum specific empirical distributions that were derived from 10K randomly drawn null (false) scores. Note that as the scores were initially perfectly calibrated the result of this step is in fact compromising the level of calibration but the goal here is to learn the effects of such 10K calibration on the of FDR estimation. While for small spectrum sets (500, 1K) the results do not differ substantially from when using perfectly calibrated scores (Figure 3), when the size of the spectrum set is comparable or larger than the 10K decoys used for the non-parametric calibration both T-TDC and mix-max slightly overestimate rather than underestimate the true FDR at small FDR values.

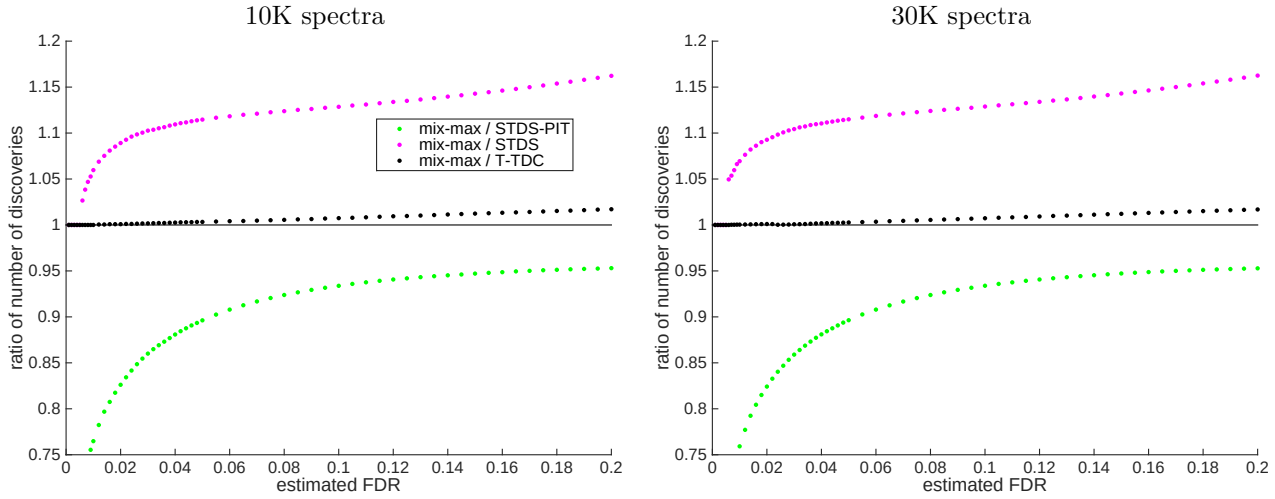

Figure 8: **Median ratios of number of discoveries, 10K calibrated scores.** Same as Figure 4 in the main text, except the data generation involved the redundant 10K calibration procedure as described in Figure 7 above. Comparing this figure with the one in the main paper we see little very little difference.

References

[1] J. D. Storey. A direct approach to false discovery rates. *J R Stat Soc Series B*, 64:479–498, 2002.